Available online at www.sciencedirect.com
ScienceDirect
Statistics & Probability Letters 77 (2007) 1034–1042
STATISTICS & PROBABILITY LETTERS
www.elsevier.com/locate/stapro

# A simple nonparametric test for diagnosing nonlinearity in Tobit median regression model

Lan Wang

*School of Statistics, University of Minnesota, 313 Ford Hall, 224 Church Street S.E., Minneapolis, Minnesota 55455, USA*

Received 20 September 2005; received in revised form 26 August 2006; accepted 17 January 2007
Available online 20 February 2007

**Abstract**

In many applications, the response variable is observed only when it is above or below a given threshold otherwise the threshold itself is observed. Tobit median regression model is a useful semiparametric procedure for analyzing this type of censored data. We propose a simple nonparametric test for assessing the common linearity assumption in this model. Compared to those existing methods in the literature, the new test has the advantage of allowing the alternative to be any smooth function. In addition, it does not require any knowledge of the parametric distribution of the random error. The test is asymptotically normal under the null hypothesis of linearity. A small Monte Carlo study demonstrates its performance.


*Keywords:* Censored data; Median regression; Nearest-neighbor windows; Quantile; Smooth alternative; Specification test; Tobit model

## 1. Introduction

Data collected from economics, epidemiology, biomedical studies and many other disciplines are often censored in the sense that the variable is not observed unless it is above or below some threshold, otherwise the threshold itself is observed. For example, in a salary survey, the data is often top coded to maintain confidentiality; in a five-year follow-up study of the life span of some cancer patients after surgery, the data is censored if the patient is still alive after five years.

The following regression model is often considered for the above type of censored data:

$$Y_i^* = m(x_i, \theta) + u_i, \tag{1.1}$$

$$Y_i = \max(Y_i^*, y_0), \tag{1.2}$$

where $(x_i, Y_i)$ are what we actually observe, $m(\cdot)$ is a function which is known up to a vector of parameters $\theta$, $u_i$'s are independent but not necessarily identically distributed random errors with median 0, $Y_i^*$ is a latent variable which is unobservable and uncensored, and $y_0$ is a given threshold. In other words, $Y_i^*$ is observed only if it exceeds $y_0$. This latent variable model is well studied and extensively applied in economics, and is

*E-mail address:* lan@stat.umn.edu.

0167-7152/$ - see front matter 
doi:10.1016/j.spl.2007.01.017

often referred to as a *Tobit* model in the econometrics literature. It is semiparametric since it assumes a parametric regression function but does not impose any parametric distribution for the random errors. The model is not restricted to left-censored data. By simple algebraic manipulation, it can easily encompass the case of right censoring (Chapter 9, Pagan and Ullah, 1999). Without loss of generality, we take $y_0 = 0$ in this paper.

The earlier work (Amemiya, 1973, Heckman, 1976, 1979) estimates the unknown parameter $\theta$ in (1.1) and (1.2) under normality assumption. These estimator are not consistent if the normality assumption does not hold. Powell (1984) proposed an estimators for this model based on least absolute deviations. His estimator is robust to misspecification of the random error distribution and permits heteroscedasticity of unknown form. The Tobit median regression model provides a valuable tool for studying the effect of a specific covariate on the possibly censored outcome and serves as a useful alternative to the traditional proportional hazards regression (Ying et al., 1995; Portnoy, 2003). Modeling the median is especially suitable for long-tailed survival data since the median can be well-estimated provided that the censoring is not too heavy while the mean may not be.

In most applications, the conditional median of the latent variable $Y^*$, i.e., $m(\cdot)$ in (1.1), is assumed to be a linear function of the covariate. This is usually for mathematical convenience. It is well known that violation of the linearity assumption can result in inconsistent estimator of the parameters and biased prediction of the survival time (Horowitz and Neumann, 1989). Classical specification tests check the linearity hypothesis by imbedding the linear function inside a larger class of parametric models, such as a class of parabolas. The classical approach is inappropriate even if the larger class of models does not fit the data well. Numerical examples in Section 4 demonstrate that the conventional large-sample Wald test can have inferior power for some nonquadratic alternatives of the conditional median. The aim of this paper is to propose a nonparametric test for diagnosing nonlinearity which permits the alternative to be any smooth function. To our knowledge, nonparametric test with smooth alternative has not been studied in this setting.

Specifically, we are interested in testing the following null hypothesis:

$$\mathrm{H}_0\colon m(x) = \beta_0 + \beta_1 x \quad \forall x, \tag{1.3}$$

for some unknown parameters $\beta_0$, $\beta_1$. The alternative functional form of $m(x)$ is allowed to be in an infinite dimensional space of smooth functions. In addition, the proposed test imposes no parametric distributional assumptions on the random error $u_i$. Specification tests for uncensored data, especially under mean zero constraint, have been well investigated, see the manuscript of Hart (1997), recent papers by Aït-Sahalia et al. (2001); Horowitz and Spokoiny (2001); He and Zhu (2003) and the references therein. In contrast, little attention has been devoted to study the same diagnostic problem for censored data. Part of the reason is that it is technically more involved.

The remainder of this paper is organized as follows. Section 2 presents the test statistic. Section 3 discusses the large sample properties. In Section 4, we report results from a Monte Carlo study. The technical proofs are given in an appendix.

## 2. Test statistic

Median regression model has the property of being equivariant to monotone transformations. As a result, if the median of $u_i$ is 0, then the conditional median of $Y_i$ given the covariate $x_i$ is $\max(0, m(x_i))$. Therefore under the null hypothesis, when the true values of $\beta_0$ and $\beta_1$ are known,

$$\varepsilon_i = I(Y_i \leqslant \max(0, \beta_0 + \beta_1 x_i)) - 1/2, \quad i = 1, \ldots, n,$$

are independent Bernoulli random variables with mean 0 and variance $\frac{1}{4}$, where $I(\cdot)$ denotes the indicator function. In practice, $\beta_0$ and $\beta_1$ have to be estimated from the data. Denote the corresponding estimators by $\widehat{\beta}_0$ and $\widehat{\beta}_1$. They are in general $\sqrt{n}$-consistent, for instance, the least absolute deviations estimator of Powell (1984) which is defined as the minimizer of $\sum_{i=1}^{n} |Y_i - \max(0, \beta_0 + \beta_1 x_i)|$.

An estimator of $\varepsilon_i$ is obtained as

$$\widehat{\varepsilon}_i = I(Y_i \leqslant \max(0, \widehat{\beta}_0 + \widehat{\beta}_1 x_i)) - 1/2.$$

We expect $\widehat{\varepsilon}_i$ to approximately have mean 0. To test for the validity of the linearity assumption (1.3), we propose to apply to $(x_i, \widehat{\varepsilon}_i)$, $i = 1, \ldots, n$, the test statistic of Wang et al. (2003), which was designed to test for constant mean regression function when the data are not censored.

The basic idea is to consider each distinct covariate as a "category" and construct a local window $W_i$ around each $x_i$, which consists of the $k_n$ nearest covariate values. This creates an artificial balanced one-way table with $n$ categories, where the responses in the $i$th category are the $\widehat{\varepsilon}_j$'s associated with the covariate values belonging to $W_i$. In what follows $W_i$ will also be understood as the set containing the indices $j$ of the covariate values that belong to the window around $x_i$, that is

$$W_i = \left\{ j : |\widehat{H}(x_j) - \widehat{H}(x_i)| \leqslant \frac{k_n - 1}{2n} \right\},$$

where $\widehat{H}(x) = n^{-1}\sum_{j=1}^{n} I\left(x_j \leqslant x\right)$. Under the null hypothesis, $\widehat{\varepsilon}_j$'s have approximately mean zero and we expect that there is approximately no treatment effects in this artificial one-way ANOVA if we think the artificial categories corresponding to different levels of a treatment. It is therefore intuitive to consider an F-type test statistic for testing no treatment effects or equivalently the test statistic MST-MSE, where MST is the treatment sum of squares and MSE is the error sum of squares, both computed from the hypothetical one-way table. This test statistic is motivated by recent development in analysis of variance with large number of factor levels (Akritas and Papadatos, 2004; Wang, 2003), where asymptotic normality of F-type test is derived for independent data when the number of factor levels goes to infinity while the number of observations per level either remains fixed or goes to infinity at appropriate rate.

This test statistic can be deemed as a generalization of the classical F-test statistic in the context of analysis of variance. Its familiar form allows quick computation by taking advantage of existing statistical software. However, its asymptotic distribution is not the same as that derived in Akritas and Papadatos (2004) or Wang (2003) because the observations in the artificial ANOVA are not independent. Different asymptotic tools have to be applied to derive the large sample theory.

Below we introduce some notations. We use $V_{ij}$, $j = 1, \ldots, k_n$, to denote the $k_n$ observations in the $i$th category, i.e., $\{V_{i1}, \ldots, V_{ik_n}\} = \{\widehat{\varepsilon}_j : j \in W_i\}$. Let $V$ be the $nk_n \times 1$ vector of all observations in this hypothetical one-way layout. The test statistic is defined as

$$T_n = \mathrm{MST} - \mathrm{MSE} = \frac{k_n}{n-1}\sum_{i=1}^{n}\left(\overline{V}_{i.} - \overline{V}_{..}\right)^2 - \frac{1}{n(k_n - 1)}\sum_{i=1}^{n}\sum_{j=1}^{k_n}\left(V_{ij} - \overline{V}_{i.}\right)^2, \tag{2.1}$$

where $\overline{V}_{i.} = k_n^{-1}\sum_{j=1}^{k_n} V_{ij}$ and $\overline{V}_{..} = n^{-1}\sum_{i=1}^{n}\overline{V}_{i.}$. Note that $T_n$ can be expressed as a quadratic form $T_n = V'AV$ with

$$A = \frac{nk_n - 1}{n(n-1)k_n(k_n - 1)}\bigoplus_{i=1}^{n} J_{k_n} - \frac{1}{n(n-1)k_n} J_{nk_n} - \frac{1}{n(k_n - 1)} I_{nk_n}, \tag{2.2}$$

where $I_{k_n}$ is the $k_n$-dimensional identity matrix, $J_{k_n} = 1_{k_n}1'_{k_n}$, $1_{k_n}$ stands for the $k_n$-dimensional column vector of 1's, and $\bigoplus$ is the notation for Kronecker (direct) sum.

In theory, other tests for testing the mean regression function which allows for smooth alternative can be similarly extended. However, this has not yet been developed, to our knowledge. The theoretical generalization from mean regression case to censored median regression is not straightforward due to the difficulty associated with the nondifferentiability of the indicator function. Usual Taylor expansion in mean regression is no longer applicable, more delicate results on uniform convergence from empirical processes are needed. An additional complication is that $\widehat{\varepsilon}_i$'s are correlated due to the estimation of the unknown parameters.

## 3. Asympototic results

Assume the design points $x_1, x_2, \ldots, x_n$ reside in a bounded interval and satisfy

$$\int_{x_i}^{x_{i+1}} r(x)\,\mathrm{d}x = \mathrm{O}(n^{-(1+s)}), \quad i = 1, \ldots, n-1, \tag{3.1}$$

for some $s > 0$, where $r(x)$ is a Lipschitz continuous positive density function. This is a regular assumption for fixed design regression. The asymptotic results presented in this section equally hold in a random design setting when $x_i$'s are independent of the errors and constitute a random sample from a continuous distribution with a positive density. Assume that the random error $u_i$ is from a continuous distribution with a positive continuous density function. It is required that the $u_i$'s have median zero, but they can have heteroscedastic distribution otherwise.

Construct a hypothetical one-way ANOVA from $(x_i, \varepsilon_i)$, $i = 1, \ldots, n$, in the way described in the previous section. Let $V^*_{ij}$, $i = 1, \ldots, n, j = 1, \ldots, k_n$, denote observations in this hypothetical one-way layout and $T^*_n$ be the corresponding test statistic computed as in (2.1). To derive the asymptotic distribution of $T_n$, we first establish the asymptotic equivalence of $T_n$ and $T^*_n$ by the following two lemmas.

**Lemma 1.** *Assume the foregoing conditions. If $n^{-1}k_n = \mathrm{o}(1)$, then under* (1.3), *as $n \to \infty$,*

$$n^{1/2}k_n^{-1/2}[V'AV - V'A_DV] \to 0$$

*in probability, where $A_D = \mathrm{diag}\,\{B_1, \ldots, B_n\}, B_i = 1/n(k_n-1)[J_{k_n} - I_{k_n}]$, $A$ is defined in (2.2).*

**Lemma 2.** *Assume the foregoing conditions. If $k_n = \mathrm{o}(n^{\gamma})$ for some $0<\gamma<1/2$, then under (1.3), as $n \to \infty$,*

$$n^{1/2}k_n^{-1/2}\left[(V^*)'A_D(V^*) - V'A_DV\right] \to 0$$

*in probability, where $A_D$ is defined as in Lemma 1.*

Since the $\varepsilon_i$'s (unobservable) are independent Bernoulli random variables with mean 0 and variance $\frac{1}{4}$, the results in Wang et al. (2003) can be used to derive the asymptotic normality of $T^*_n$. This leads to the main theorem.

**Theorem 1.** *Assume the foregoing conditions. If $k_n \to \infty$, $k_n = \mathrm{o}(n^{\gamma})$, for some $0<\gamma<1/2$, then under (1.3), as $n \to \infty$,*

$$n^{1/2}k_n^{-1/2}T_n \to \mathrm{N}(0, 1/12),$$

*in distribution.*

As a result of this theorem, a test rejects $H_0$ when $n^{1/2}k_n^{-1/2}T_n > Z_{1-\alpha}/\sqrt{12}$, where $Z_{1-\alpha}$ is the $1-\alpha$ quantile of a standard normal distribution, has asymptotic level $\alpha$. It is also possible to have $k_n = k$, a fixed constant. In this case, we have under (1.3), as $n \to \infty$, $n^{1/2}T_n \to \mathrm{N}(0, [24(k-1)]^{-1}k(2k-1))$ in distribution. For any given fixed alternative $m(x, \theta)$ such that $\int (P(Y < \max(0, \beta^*_0 + \beta^*_1 x)) - 1/2)^2 r(x)\,\mathrm{d}x - (\int (P(Y < \max(0, \beta^*_0 + \beta^*_1 x)) - \frac{1}{2})r(x)\,\mathrm{d}x)^2 \neq 0$ where $Y = \max(0, m(x, \theta) + u)$ with $u$ having median 0 and $(\beta^*_0, \beta^*_1)$ are the estimated coefficients under the null hypothesis, we can show that $n^{1/2}k_n^{-1/2}T_n \to \infty$ in probability. Therefore, the test is consistent against a large class of alternatives. We also conjecture that the proposed test can detect local alternative that converges to the null at the rate $(nk_n)^{-1/4}$, similarly as in the setting of testing for the validity of mean regression model.

The testing procedure can be easily adapted to handle general censored quantile regression. If $P(u_i < 0) = \tau$, $0 < \tau < 1$, then the pseudo-residual can be taken as $\varepsilon_i = I(Y_i \leqslant \max(0, \beta_0 + \beta_1 x_i)) - \tau$, the test statistic and the asymptotic theory can be developed similarly.

## 4. A Monte Carlo Study

To investigate the finite-sample size and power performance of the proposed specification test, results from a small Monte Carlo study is reported in this section. For comparison purpose, we also include results from the classical asymptotic Wald test for censored median regression. For the Wald test to assess linearity, a quadratic model is fitted first and the null hypothesis of linearity is rejected if the quadratic coefficient is significantly different from zero.

The following censored median regression model is under consideration:

$$Y_i = \max(0, 0.6 + \phi(X_i) + \varepsilon_i), \quad i = 1, \ldots, 100, \tag{4.1}$$

where $X_i$ is uniformly distributed on $(-1, 1)$, $\varepsilon_i$ has standard normal distribution. It is of interest to test the null hypothesis (1.3). Several different types of alternatives for $\phi(x)$ are considered, including:

$$\phi(x) = x + ax^2, \quad a = 0, 1, 2, 3,$$
$$\phi(x) = x + b(\sin(2\pi x))^2, \quad b = 1, 2, 3,$$
$$\phi(x) = c(x\sin(2\pi x))^2, \quad c = 2, 3, 4.$$

The case $a = 0$ corresponds to the null hypothesis, under which the proportion of censoring is about 30%. The random data are generated using the software Matlab. In 500 simulation runs, the proportion of rejection is summarized in Table 1 for different local window size $k_n = 9, 11, 13$. It is observed that the empirical level of the proposed test is quite close to the nominal level $\alpha = 0.05$. When the true model is a quadratic model, the Wald test is the most powerful test. This is the price the nonparametric test needs to pay in order to be omnibus. For the other two types of alternatives, the superiority of the new test is self-evident.

To assess the effect of heteroscedastic errors, we repeat the above simulation with the i.i.d random error $\varepsilon_i$ replaced by the heteroscedastic random error $\sigma(X_i)\varepsilon_i$, where $\sigma^2(x) = 0.75(1 + x^2)$ is normalized to satisfy $\int \sigma^2(x)r(x)\,\mathrm{d}x = 1$ with $r(x)$ denoting the density of the covariate. The results are summarized in Table 2. It is observed that the nonparametric test maintains accurate level under the heteroscedastic random error. Compared with the homoscedastic error case, the observed power exhibits slight change whose direction and magnitude depends on the specific alternative. For instance, for the quadratic alternative with $a = 1$, the power for heteroscedastic case is slightly lower than that for homoscedastic case.

## 5. Discussions

The test statistic discussed in the current paper can be generalized to multiple regression case. This requires construction of artificial higher-way ANOVA, see for instance Wang (2003) for two-way ANOVA with large number of levels, and becomes more computationally intensive. A new lack-of-fit test based on kernel method is under study where multiple covariates can be more easily incorporated.

Table 1
Observed level and power at nominal level 5% for the censored median regression model: $Y_i = \max(0, 0.6 + \phi(X_i) + \mathrm{N}(0, 1))$, $i = 1, \ldots, 100$

| $\phi(x)$ | Nonparametric test | | | | |
|---|---|---|---|---|---|
| | | $k_n = 9$ | $k_n = 11$ | $k_n = 13$ | Wald test |
| $x + ax^2$ | a = 0 | 0.058 | 0.050 | 0.052 | 0.046 |
| | a = 1 | 0.170 | 0.174 | 0.172 | 0.560 |
| | a = 2 | 0.752 | 0.740 | 0.704 | 0.968 |
| | a = 3 | 0.986 | 0.986 | 0.980 | 1.000 |
| $x + b(\sin(2\pi x))^2$ | b = 1 | 0.166 | 0.118 | 0.090 | 0.068 |
| | b = 2 | 0.792 | 0.702 | 0.530 | 0.076 |
| | b = 3 | 0.998 | 0.996 | 0.952 | 0.086 |
| $c(x\,\sin(2\pi x))^2$ | c = 2 | 0.368 | 0.356 | 0.336 | 0.374 |
| | c = 3 | 0.728 | 0.716 | 0.686 | 0.588 |
| | c = 4 | 0.942 | 0.936 | 0.914 | 0.754 |

Table 2
Observed level and power at nominal level 5% for the censored median regression model: $Y_i = \max(0, 0.6 + \phi(X_i) + \sigma(X_i)N(0,1))$, where $\sigma^2(X_i) = 0.75(1 + X_i^2)$, $i = 1, \ldots, 100$

| $\phi(x)$ | Nonparametric test | | | | |
|---|---|---|---|---|---|
| | | $k_n = 9$ | $k_n = 11$ | $k_n = 13$ | Wald test |
| $x + ax^2$ | a = 0 | 0.058 | 0.058 | 0.056 | 0.062 |
| | a = 1 | 0.164 | 0.156 | 0.148 | 0.560 |
| | a = 2 | 0.738 | 0.722 | 0.718 | 0.966 |
| | a = 3 | 0.980 | 0.980 | 0.976 | 1.000 |
| $x + b(\sin(2\pi x))^2$ | b = 1 | 0.200 | 0.148 | 0.104 | 0.058 |
| | b = 2 | 0.856 | 0.786 | 0.606 | 0.078 |
| | b = 3 | 1.000 | 0.994 | 0.972 | 0.082 |
| $c(x\ \sin(2\pi x))^2$ | c = 2 | 0.350 | 0.326 | 0.308 | 0.410 |
| | c = 3 | 0.742 | 0.730 | 0.684 | 0.636 |
| | c = 4 | 0.932 | 0.922 | 0.904 | 0.774 |

A problem that has not been solved in this paper is how to choose $k_n$. This is in fact an open problem facing all smoothing-based tests. In general, the optimal smoothing parameter for hypothesis testing is not the same as that for curve estimation, see Section 6.4 of Hart (1997) for related discussions. For any practical problem, King et al. (1991), Young and Bowman (1995) suggest calculating $p$-values for a wide range choices of the smoothing parameter, and called the plot of $P$-values versus the smoothing parameter a "significant trace". An alternative method is proposed by Horowitz and Spokoiny (2001) to circumvent the problem of choosing a single smoothing parameter. Their idea is to define a new test statistic that is taken as the maximum of the original test over a sequence of smoothing parameters. They prove that this leads to a test with certain optimality property. However, the maximal test has a different asymptotic distribution and the derivation of its theoretical property is quite involved.

## Appendix A

### A.1. Proofs

Let $\beta = (\beta_0, \beta_1)'$, $\widehat{\beta} = (\widehat{\beta}_0, \widehat{\beta}_1)'$, $\beta^* = (\beta_0^*, \beta_1^*)'$, $\widetilde{\beta} = (\widetilde{\beta}_0, \widetilde{\beta}_1)'$ and let $||\cdot||$ denote the usual Euclidean norm for vectors. Define $\varepsilon_j(\beta) = I(Y_j \leqslant \max(0, \beta_0 + \beta_1 x_j)) - \frac{1}{2}$ and $Z_j(\beta^*, \beta) = \varepsilon_j(\beta^*) - \varepsilon_j(\beta)$.

We first state and prove a useful probability bound.

**Lemma A1.** *If* $||\beta - \beta^*|| = O(n^{-\alpha})$, *for some* $0 < \alpha \leqslant 1/2$, *then*

$$\sum_{i=1}^{n} |Z_i(\beta^*, \beta)| = O_p(n^{1-\alpha}).$$

**Proof.** We can show that

$$|Z_j(\beta^*, \beta)| \leqslant I(|u_j| \leqslant |\beta_0^* - \beta_0| + |x_j||\beta_1^* - \beta_1|) \leqslant I(|u_j| \leqslant c^*||\beta^* - \beta||),$$

for some positive constant $c^*$. Since $\xi_i = I(|u_i| \leqslant c^*||\beta^* - \beta||)$ are independent Bernoulli random variables with mean and variance both of order $O(n^{-\alpha})$, by Bernstein's inequality, we have for any $t > 0$, $\forall\ n$,

$$P\left(\left|\sum_{j=1}^{n}(\xi_j - E(\xi_j))\right| > t\right) \leqslant 2e^{-t^2/2\left(\sum_{j=1}^{n} Var(\xi_j) + t\right)}.$$

Take $t = n^{1-\alpha}$, using the fact $\sum_{j=1}^{n} E(\xi_j) = \sum_{j=1}^{n} Var(\xi_j) = O(n^{1-\alpha})$, we have there exist $M > 0$ and a positive integer $n_0$ such that for $\forall\ n > n_0$,

$$P\left(\sum_{j=1}^{n} \xi_j > Mn^{1-\alpha}\right) \leqslant 2e^{-c'n^{1-\alpha}}.$$

for some $c' > 0$. Thus $\sum_{i=1}^{n} |Z_i(\beta^*, \beta)| \leqslant \sum_{j=1}^{n} \xi_j = O_p(n^{1-\alpha})$. $\square$

**Proof of Lemma 1.** Note that

$$\begin{aligned}
&n^{1/2}k_n^{-1/2}[V'AV - V'A_D V] \\
&= -n^{1/2}k_n^{-1/2}[n(n-1)k_n]^{-1} \sum_{i_1 \neq i_2}^{n} \sum_{j_1=1}^{n} \sum_{j_2=1}^{n} \widehat{\varepsilon}_{j_1} \widehat{\varepsilon}_{j_2} I(j_1 \in W_{i_1}, j_2 \in W_{i_2}) \\
&= O(n^{-3/2}k_n^{-3/2}) \sum_{i_1 \neq i_2}^{n} \sum_{j_1=1}^{n} \sum_{j_2=1}^{n} \varepsilon_{j_1}(\widehat{\beta}) \varepsilon_{j_2}(\widehat{\beta}) I(j_1 \in W_{i_1}, j_2 \in W_{i_2}).
\end{aligned}$$

Define

$$Q(\beta^*) = O(n^{-3/2}k_n^{-3/2}) \sum_{i_1 \neq i_2}^{n} \sum_{j_1=1}^{n} \sum_{j_2=1}^{n} \varepsilon_{j_1}(\beta^*) \varepsilon_{j_2}(\beta^*) I(j_1 \in W_{i_1}, j_2 \in W_{i_2}).$$

Since $||\widehat{\beta} - \beta|| = O_p(n^{-1/2})$, to prove the lemma, it is sufficient to show that for any $c > 0$,

$$\sup_{||\beta^* - \beta|| \leqslant cn^{-1/2}} |Q(\beta^*)| = o_p(1).$$

$Q(\beta^*)$ can be decomposed as: $Q(\beta^*) = Q_1(\beta^*, \beta) + Q_2(\beta) + Q_3(\beta^*, \beta)$, where

$$\begin{aligned}
Q_1(\beta^*, \beta) &= O(n^{-3/2}k_n^{-3/2}) \sum_{i_1 \neq i_2}^{n} \sum_{j_1=1}^{n} \sum_{j_2=1}^{n} Z_{j_1}(\beta^*, \beta) Z_{j_2}(\beta^*, \beta) I(j_1 \in W_{i_1}, j_2 \in W_{i_2}), \\
Q_2(\beta) &= O(n^{-3/2}k_n^{-3/2}) \sum_{i_1 \neq i_2}^{n} \sum_{j_1=1}^{n} \sum_{j_2=1}^{n} \varepsilon_{j_1}(\beta) \varepsilon_{j_2}(\beta) I(j_1 \in W_{i_1}, j_2 \in W_{i_2}), \\
Q_3(\beta^*, \beta) &= O(n^{-3/2}k_n^{-3/2}) \sum_{i_1 \neq i_2}^{n} \sum_{j_1=1}^{n} \sum_{j_2=1}^{n} Z_{j_1}(\beta^*, \beta) \varepsilon_{j_2}(\beta) I(j_1 \in W_{i_1}, j_2 \in W_{i_2}).
\end{aligned}$$

By directly checking the mean and the variance, we have $Q_2(\beta) = o_p(1)$. We have

$$|Q_1(\beta^*, \beta)| \leqslant O(n^{-3/2}k_n^{-3/2}) \left[\sum_{i=1}^{n} \sum_{j=1}^{n} |Z_j(\beta^*, \beta)| I(j \in W_i)\right]^2$$

By Lemma A1 (with $\alpha = \frac{1}{2}$),

$$\sup_{||\beta^* - \beta|| \leqslant cn^{-1/2}} |Q_1(\beta^*, \beta)| \leqslant O(n^{-3/2}k_n^{-3/2}) O_p(nk_n^2) = O_p(n^{-1/2}k_n^{1/2}) = o_p(1).$$

Similarly,

$$\begin{aligned}
&\sup_{||\beta^* - \beta|| \leqslant cn^{-1/2}} |Q_3(\beta^*, \beta)| \\
&\leqslant O(n^{-3/2}k_n^{-3/2}) \left|\sum_{i_1=1}^{n} \sum_{j_1=1}^{n} Z_{j_1}(\beta^*, \beta) I(j_1 \in W_{i_1})\right| \left|\sum_{i_2=1}^{n} \sum_{j_2=1}^{n} \varepsilon_{j_2}(\beta) I(j_2 \in W_{i_2})\right| \\
&= O(n^{-3/2}k_n^{-3/2}) O_p(n^{1/2}k_n) O_p(n^{1/2}k_n) = o_p(1).
\end{aligned}$$

So Lemma 1 is proved. $\square$

**Proof of Lemma 2.** Note that

$$\begin{aligned}
&n^{1/2}k_n^{-1/2}[(V^*)'A_D(V^*) - V'A_DV] \\
&= n^{1/2}k_n^{-1/2}[n(k_n-1)]^{-1}\sum_{i=1}^{n}\sum_{j_1\neq j_2}^{n}(\varepsilon_{j_1}\varepsilon_{j_2} - \widehat{\varepsilon}_{j_1}\widehat{\varepsilon}_{j_2})I(j_1,j_2 \in W_i) \\
&= n^{-1/2}k_n^{-1/2}(k_n-1)^{-1}\sum_{i=1}^{n}\sum_{j_1\neq j_2}^{n}\varepsilon_{j_1}(\varepsilon_{j_2} - \widehat{\varepsilon}_{j_2})I(j_1,j_2 \in W_i) \\
&\quad + n^{-1/2}k_n^{-1/2}(k_n-1)^{-1}\sum_{i=1}^{n}\sum_{j_1\neq j_2}^{n}\widehat{\varepsilon}_{j_2}(\varepsilon_{j_1} - \widehat{\varepsilon}_{j_1})I(j_1,j_2 \in W_i) \\
&= D_1 + D_2,
\end{aligned}$$

where the definition of $D_1$ and $D_2$ should be clear from the context. Define

$$D_1(\beta^*,\beta) = n^{-1/2}k_n^{-1/2}(k_n-1)^{-1}\sum_{i=1}^{n}\sum_{j_1\neq j_2}^{n}\varepsilon_{j_1}(\beta)Z_{j_2}(\beta^*,\beta)I(j_1,j_2 \in W_i).$$

We shall show that $\sup_{||\beta^*-\beta||\leqslant cn^{-1/2}}|D_1(\beta^*,\beta)| = o_p(1)$ for any given constant $c>0$. It is useful to first derive the mean and variance of $D_1(\beta^*,\beta)$. First, $ED_1(\beta^*,\beta)=0$. A tedious but otherwise routine calculation yields $ED_1^2(\beta^*,\beta) = O(n^{-1}k_n) + O(n^{-1/2})$ for $||\beta^*-\beta||\leqslant cn^{-1/2}$.

By assumption, there exists $0<\gamma<1/2$ such that $k_n = o(n^{\gamma})$. Take $\alpha=(1+\gamma)/2$, then $k_n = o(n^{2\alpha-1})$. We cover the ball $\{\beta^* : ||\beta^*-\beta||\leqslant cn^{-1/2}\}$ with a net of balls, each of them has radius $n^{-\alpha}$. This net can be constructed with cardinality $l_n = O(n^{\gamma}) = O(n^{-1+2\alpha})$. Let $D_n = \{\widetilde{\beta}_1,\ldots,\widetilde{\beta}_{l_n}\}$ be the collection of the centers of the balls in the cover, denote $B(\widetilde{\beta}_j)$ the ball with center $\widetilde{\beta}_j$, then

$$\begin{aligned}
P\left(\sup_{||\beta^*-\beta||\leqslant cn^{-1/2}}|D_1(\beta^*,\beta)|>\varepsilon\right) &\leqslant \sum_{j:\widetilde{\beta}_j\in D_n} P\left(\sup_{\beta^*\in B(\widetilde{\beta}_j)}|D_1(\beta^*,\beta)|>\varepsilon\right) \\
&\leqslant \sum_{j:\widetilde{\beta}_j\in D_n} P\left(\sup_{||\beta^*-\widetilde{\beta}_j||\leqslant O(n^{-\alpha})}|D_1(\beta^*,\widetilde{\beta}_j)|>\varepsilon/2\right) \\
&\quad + \sum_{j:\widetilde{\beta}_j\in D_n} P(|D_1(\widetilde{\beta}_j,\beta)|>\varepsilon/2),
\end{aligned} \tag{A.1}$$

where the last step is obtained by triangular inequality. By Chebyshev's inequality, the second term of (A.1) is bounded by

$$4l_n\frac{ED_1^2(\widetilde{\beta}_j,\beta)}{\varepsilon^2} = 4l_n\frac{O(n^{-1}k_n)+O(n^{-1/2})}{\varepsilon^2} = o(n^{-1+2\gamma}) + O(n^{-1/2+\gamma}) = o(1).$$

To bound the first term, first notice that

$$\begin{aligned}
&|D_1(\beta^*,\widetilde{\beta}_j)| \\
&\leqslant n^{-1/2}k_n^{-1/2}(k_n-1)^{-1}\sum_{i=1}^{n}\sum_{j_1\neq j_2}^{n} I(|u_{j_2}|\leqslant|\widetilde{\beta}_{j_2 0}-\beta_0^*| + |x_{j_2}||\widetilde{\beta}_{j_2 1}-\beta_1^*|)I(j_1,j_2\in W_i) \\
&\leqslant 2n^{-1/2}k_n^{1/2}\sum_{j=1}^{n} I(|u_j|\leqslant c^*||\beta^*-\widetilde{\beta}_j||),
\end{aligned}$$

for some $c^* > 0$. Therefore the first term of (A.1) is bounded by

$$\sum_{j:\widetilde{\beta}_j \in D_n} P\left(\sum_{j=1}^{n} I(|u_j| \leqslant c^* ||\beta^* - \widetilde{\beta}_j||) > \frac{\varepsilon}{4} k_n^{-1/2} n^{1/2}\right). \tag{A.2}$$

If $||\beta^* - \widetilde{\beta}_j|| \leqslant O(n^{-\alpha})$, by Bernstein's inequality applied to the Bernoulli random variables $\xi_j = I(|u_j| \leqslant c^* ||\beta^* - \widetilde{\beta}_j||)$ and use the fact $\sum_{i=1}^{n} E(\xi) = \sum_{j=1}^{n} Var(\xi) = \mathrm{O}(n^{1-\alpha})$ we have there exits $c_1 > 0$ and $c_2 > 0$ and a positive integer $n_0$ such that $\forall n > n_0$, $\forall t > 0$,

$$P\left(\sum_{j=1}^{n} I(|u_j| \leqslant c^* ||\beta^* - \beta||) > t + c_1 n^{1-\alpha}\right) \leqslant 2\mathrm{e}^{-t^2/2(c_2 n^{1-\alpha} + t)}.$$

Take $t = \varepsilon/4 n^{1/2} k_n^{-1/2} - c_1 n^{1-\alpha}$, using the fact $n^{1-\alpha} = \mathrm{o}(n^{1/2} k_n^{-1/2})$, the above inequality becomes

$$P\left(\sum_{j=1}^{n} I(|u_j| \leqslant c^* ||\beta^* - \widetilde{\beta}_j||) > \frac{\varepsilon}{4} n^{1/2} k_n^{-1/2}\right) \leqslant 2\mathrm{e}^{-c_3 n^{1/2} k_n^{-1/2}},$$

for some $c_3 > 0$ and $n > n_0$. Therefore (A.2) is bounded by $\mathrm{O}(n^{\gamma})\mathrm{O}(\mathrm{e}^{-3 n^{1/2} k_n^{-1/2}}) = \mathrm{o}(1)$. Combining the above derivation, $D_1(\beta^*, \beta) = \mathrm{o_p}(1)$. Similarly, we can show $D_2 = \mathrm{o_p}(1)$. □

**Proof of Theorem 1.** Since the $\varepsilon_i$'s are independent Bernoulli random variables with mean 0 and variance $\frac{1}{4}$, we apply the results of Wang et al. (2003) and obtain

$$n^{1/2} k_n^{-1/2} (V^*)' A_D (V^*) \xrightarrow{\mathrm{d}} N(0, 1/12).$$

The proof is done by combining the results from the two lemmas. □

## References


Aït-Sahalia, Bickel, P.J., Stoker, T.M., 2001. Goodness-of-fit tests for kernel regression with an application to option implied volatilities. J. Econometrics 105, 363–412.

Akritas, M.G., Papadatos, N., 2004. Heteroscedastic one-way ANOVA and lack-of-fit test. J. Amer. Statist. Assoc. 99, 368–382.

Amemiya, T., 1973. Regression analysis when the dependent variable is truncated normal. Econometrica 41, 997–1016.

Hart, J., 1997. Nonparametric Smoothing and Lack-of-fit Test. Springer, New York.

He, X., Zhu, L.X., 2003. A lack-of-fit test for quantile regression. J. Amer. Statist. Assoc. 98, 1013–1022.

Heckman, J., 1976. The common structure of statistical models of truncation, sample selection, and limited dependent variables and a simple estimator for such models. Ann. Econom. Social Meas. 5, 475–492.

Heckman, J., 1979. Sample bias as a specification error. Econometrica 47, 153–162.

Horowitz, J.L., Neumann, G.R., 1989. Specification testing in censored regression models: parametric and semiparametric methods. J. Appl. Econometrics 4, S61–86.

Horowitz, J.L., Spokoiny, V.G., 2001. An adaptive, rate-optimal test of a parametric mean-regression model against a nonparametric alternative. Econometrica 69, 599–631.

Pagan, A., Ullah, A., 1999. Nonparametric Econometrics. Cambridge University Press, Cambridge, UK.

Portnoy, S., 2003. Censored regression quantiles. J. Amer. Statis. Assoc. 98, 1001–1012.

Powell, J.L., 1984. Least absolue deviations estimation for the censored regression model. J. Econometrics 25, 303–325.

Wang, L., 2003. Heteroscedastic ANOVA with large number of factor levels and nonparametric ANCOVA. Ph.D. thesis. The Pennsylvania State University.

Wang, L., Akritas, M.G., Van Keilegom, I., 2003. Nonparametric Goodness-of-fit test for heteroscedastic regression models. Submitted for publication.

Ying, Z., Jung, S.H., Wei, L.J., 1995. Survival analysis with median regression models. J. Amer. Statist. Assoc. 90, 178–184.